Online Social Networks and Media

Network Measurements and Models
Measuring and Modeling Networks

• There are networks everywhere
• What do they look like?
  – How do you measure and describe a billion node network?
• What are the process that generate them?
  – Can we create models for real-life networks?
• These two questions are related: We need to measure the characteristics that we want to model
Before we start

• Wait, there is a model for generating graphs!
• The Erdős-Renyi $G_{n,p}$ random graph model:
  – $n$: the number of vertices
  – $p$: probability of generating an edge
    • for each pair $(i,j)$, generate the edge $(i,j)$ independently with probability $p$
• A very well studied model in graph theory!
  – As we will see, not good enough in our case
Measuring Networks

- Degree distributions and power-laws
- Clustering Coefficient
- Small world phenomena
- Components
- Motifs
- Homophily
Degree distributions

\[ f_k = \text{fraction of nodes with degree } k \]
\[ = \text{probability of a randomly selected node to have degree } k \]
It all started with some Greeks


![Graphs showing degree distributions for the internet graph](image)

Figure 6: The outdegree plots: Log-log plot of frequency $f_d$ versus the outdegree $d$.

- Degree distributions for the internet graph
Power-law distributions

• The degree distributions of most real-life networks follow a power law
  \[ p(k) = Ck^{-\alpha} \]

• Right-skewed/Heavy-tail distribution
  – there is a non-negligible fraction of nodes that has very high degree (hubs)
  – scale-free: no characteristic scale, average is not informative

• In stark contrast with the random graph model!
  – Poisson degree distribution, \( z=np \)
    \[ p(k) = \frac{z^k}{k!} e^{-z} \]
    – Concentrated around the mean
    – the probability of very high degree nodes is exponentially small
Power-law signature

- Power-law distribution gives a line in the log-log plot

\[ \log p(k) = -\alpha \log k + \log C \]

- \( \alpha \): power-law exponent (typically \( 2 \leq \alpha \leq 3 \))
A random graph example
Power-laws appear in all networks!

Taken from [Newman 2003]
And not only in networks!

<table>
<thead>
<tr>
<th>quantity</th>
<th>$x_{\text{min}}$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency of use of words</td>
<td>1</td>
<td>2.20(1)</td>
</tr>
<tr>
<td>number of citations to papers</td>
<td>100</td>
<td>3.04(2)</td>
</tr>
<tr>
<td>number of hits on web sites</td>
<td>1</td>
<td>2.40(1)</td>
</tr>
<tr>
<td>copies of books sold in the US</td>
<td>2000000</td>
<td>3.51(16)</td>
</tr>
<tr>
<td>telephone calls received</td>
<td>10</td>
<td>2.22(1)</td>
</tr>
<tr>
<td>magnitude of earthquakes</td>
<td>3.8</td>
<td>3.04(4)</td>
</tr>
<tr>
<td>diameter of moon craters</td>
<td>0.01</td>
<td>3.14(5)</td>
</tr>
<tr>
<td>intensity of solar flares</td>
<td>200</td>
<td>1.83(2)</td>
</tr>
<tr>
<td>intensity of wars</td>
<td>3</td>
<td>1.80(9)</td>
</tr>
<tr>
<td>net worth of Americans</td>
<td>$600\text{m}$</td>
<td>2.09(4)</td>
</tr>
<tr>
<td>frequency of family names</td>
<td>100000</td>
<td>1.94(1)</td>
</tr>
<tr>
<td>population of US cities</td>
<td>400000</td>
<td>2.30(5)</td>
</tr>
</tbody>
</table>

TABLE I Parameters for the distributions shown in Fig. 4. The labels on the left refer to the panels in the figure. Exponent values were calculated using the maximum likelihood method of Eq. (5) and Appendix B, except for the moon craters (g), for which only cumulative data were available. For this case the exponent quoted is from a simple least-squares fit and should be treated with caution. Numbers in parentheses give the standard error on the trailing figures.
Measuring power-laws

• How do we create these plots? How do we measure the power-law exponent?

• Collect a set of measurements:
  – E.g., the degree of each page, the number of appearances of each word in a document, the size of solar flares (continuous)

• Create a value histogram
  – For discrete values, this consists of pairs of the form (value, number of times the value appears)
  – For continuous values (but also for discrete):
    • Break the range of values into bins of equal width
    • Sum the count of values in the bin
    • Represent the bin by the mean (median) value
  – The histogram consists of pairs (mean bin value, count of values in bin)

• Plot the pairs in log-log scale
Discrete Counts

Word Count Plot
Measuring power laws

Simple binning produces a noisy plot
Logarithmic binning

• **Exponential binning**
  – Create bins that grow *exponentially* in size
  – In each bin divide the **count** by the **bin length**
  • This is the number of **observations per bin unit**

Still some noise at the tail
Cumulative distribution

• Compute the cumulative distribution
  – $\Pr[X \geq x]$: fraction (or number) of observations that have value at least $x$
  – It also follows a power-law with exponent $\alpha - 1$
Pareto distribution

• A random variable follows a Pareto distribution if

\[ P[X \geq x] = C' x^{-\beta} \quad \text{for} \quad x \geq x_{\text{min}} \]

• Power law distribution with exponent \( \alpha = 1 + \beta \)
There is another easy way to see the power-law, by doing the Zipf plot

- Order the values in decreasing order
- Plot the values against their rank in log-log scale
  - i.e., for the r-th value $x_r$, plot the point $(\log(r), \log(x_r))$
- If there is a power-law you should see something like a straight line
Zipf’s Law

• A random variable $X$ follows Zipf’s law if the $r$-th largest value $x_r$ satisfies
  \[ x_r \approx r^{-\gamma} \]

• Same as Pareto distribution
  \[ P[X \geq x] \approx x^{-1/\gamma} \]

• $X$ follows a power-law distribution with $\alpha=1+1/\gamma$

• Named after Zipf, who studied the distribution of words in English language and found Zipf law with exponent 1
Zipf vs Pareto
Computing the exponent

• Maximum likelihood estimation
  – Assume that the set of data observations $x$ are produced by a power-law distribution with some exponent $\alpha$
    • Exact law: $p(x) = \frac{\alpha-1}{x_{\text{min}}} \left( \frac{x}{x_{\text{min}}} \right)^{-\alpha}$
  – Find the exponent that maximizes the probability $P(\alpha \mid x)$
    $$\alpha = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}} \right]^{-1}$$

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>( n )</th>
<th>( m )</th>
<th>( z )</th>
<th>( \ell )</th>
<th>( \alpha )</th>
<th>( C^{(1)} )</th>
<th>( C^{(2)} )</th>
<th>( r )</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20,416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>-</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105,323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>-</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107,182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311,313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>-</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311,313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>-</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>-</td>
<td>3.2</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>14,34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>351</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5103</td>
<td>4.99</td>
<td>4.87</td>
<td>-</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td></td>
<td></td>
<td>119,157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
<td>86,148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>-</td>
<td>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>-</td>
<td>0.69</td>
<td>-0.033</td>
<td>0.366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>-0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>-</td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,097</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>-0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
<td>6,354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>-</td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>-</td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>-</td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416,421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices \( n \); total number of edges \( m \); mean degree \( z \); mean vertex–vertex distance \( \ell \); exponent \( \alpha \) of degree distribution if the distribution follows a power law (or \( \alpha = \infty \) if not; in/out-degree exponents are given for directed graphs); clustering coefficient \( C^{(1)} \) from Eq. (3); clustering coefficient \( C^{(2)} \) from Eq. (6); and degree correlation coefficient \( r \), Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Power Laws - Recap

- A (continuous) random variable $X$ follows a power-law distribution if it has density function
  \[ p(x) = Cx^{-\alpha} \]

- A (continuous) random variable $X$ follows a Pareto distribution if it has cumulative function
  \[ P[X \geq x] = Cx^{-\beta} \quad \text{power-law with } \alpha=1+\beta \]

- A (discrete) random variable $X$ follows Zipf’s law if the $r$-th largest value satisfies
  \[ x_r = Cr^{-\gamma} \quad \text{power-law with } \alpha=1+1/\gamma \]
Average/Expected degree

• For power-law distributed degree
  – if $\alpha \geq 2$, it is a constant
    \[ E[X] = \frac{\alpha - 1}{\alpha - 2} x_{\text{min}} \]
  – if $\alpha < 2$, it diverges
    • The expected value goes to infinity as the size of the network grows

• The fact that $\alpha \geq 2$ for most real networks guarantees a constant average degree as the graph grows
The 80/20 rule

- **Top-heavy**: Small fraction of values collect most of distribution mass

  - This phenomenon becomes more extreme when $\alpha < 2$
  - 1% of values has 99% of mass
  - E.g. name distribution
The effect of exponent

As the exponent increases the probability of observing an extreme value decreases.
Clustering (Transitivity) coefficient

- Measures the density of triangles (local clusters) in the graph
- Two different ways to measure it:
  \[ C^{(1)} = \frac{\sum_{i} \text{triangles centered at node } i}{\sum_{i} \text{triples centered at node } i} \]
- The ratio of the means
Example

\[ C^{(1)} = \frac{3}{1 + 1 + 6} = \frac{3}{8} \]
Clustering (Transitivity) coefficient

- Clustering coefficient for node $i$

\[
C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i}
\]

\[
C^{(2)} = \frac{1}{n} C_i
\]

- The mean of the ratios
- This is the clustering coefficient that we will use.
Example

The two clustering coefficients give different measures

$C^{(2)} = \frac{1}{5}(1 + 1 + 1/6) = \frac{13}{30}$

$C^{(1)} = \frac{3}{8}$

- The two clustering coefficients give different measures
- $C^{(2)}$ increases with nodes with low degree
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>( n )</th>
<th>( m )</th>
<th>( z )</th>
<th>( \ell )</th>
<th>( \alpha )</th>
<th>( C^{(1)} )</th>
<th>( C^{(2)} )</th>
<th>( r )</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>8,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>2.1</td>
<td>0.16</td>
<td></td>
<td></td>
<td>8, 9</td>
</tr>
<tr>
<td>email messages</td>
<td>direct</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>direct</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.61</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>direct</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>direct</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td></td>
<td></td>
<td></td>
<td>74</td>
</tr>
<tr>
<td>citation network</td>
<td>direct</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/5.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>351</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>direct</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td></td>
<td></td>
<td>119, 157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.09</td>
<td>0.033</td>
<td>–0.366</td>
<td>366</td>
</tr>
<tr>
<td>software packages</td>
<td>direct</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>direct</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.038</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>direct</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>direct</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>direct</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a network of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices \( n \); total number of edges \( m \); mean degree \( z \); mean vertex-vertex distance \( \ell \); exponent \( \alpha \) of degree distribution if the distribution follows a power law (or log if not; in/out-degree exponents are given for directed graphs); clustering coefficient \( C^{(1)} \) from Eq. (3); clustering coefficient \( C^{(2)} \) from Eq. (6); and degree correlation coefficient \( r \), Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Clustering coefficient for random graphs

- The probability of two of your neighbors also being neighbors is $p$, independent of local structure
  - clustering coefficient $C = p$
  - when the average degree $z=np$ is constant $C = O(1/n)$

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$z$</th>
<th>$C$ measured</th>
<th>$C$ for random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet [153]</td>
<td>6,374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World Wide Web (sites) [2]</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>power grid [192]</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>biology collaborations [140]</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>mathematics collaborations [141]</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>film actor collaborations [149]</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>company directors [149]</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>word co-occurrence [90]</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>neural network [192]</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>metabolic network [69]</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>food web [138]</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Small worlds

- **Millgram’s experiment**: Letters were handed out to people in Nebraska to be sent to a target in Boston.
- People were instructed to pass on the letters to someone they knew on first-name basis.
- The letters that reached the destination followed paths of length around 6.
- **Six degrees of separation**: (play of John Guare)

- Also:
  - The Kevin Bacon game
  - The Erdös number
Measuring the small world phenomenon

- $d_{ij} =$ **shortest path** between $i$ and $j$

- **Diameter:**
  
  $$d = \max_{i,j} d_{ij}$$

- **Characteristic path length:**
  
  $$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- **Harmonic mean**
  
  $$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

- Also, distribution of all shortest paths
Effective Diameter

- Disconnected components or isolated long paths can throw off the computation of the diameter.
- **Effective diameter**: the interpolated value where 90% of node pairs are reachable.

**Computation:**
- \( f(d) \): for integer \( d \), the fraction of pairs in the graph that have distance less or equal to \( D \).
- \( f(x) \): for real \( x \): \( d - 1 < x < d \), \( f(x) = \frac{f(d) - f(d-1)}{x-d} \).
- **Effective Diameter**: the real value \( x \) such that \( f(x) = 0.9 \).
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>( z )</th>
<th>( \ell )</th>
<th>( \alpha )</th>
<th>( C^{(1)} )</th>
<th>( C^{(2)} )</th>
<th>( r )</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449913</td>
<td>25516482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253339</td>
<td>496489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52909</td>
<td>245300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1520251</td>
<td>11803064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47000000</td>
<td>80000000</td>
<td>3.16</td>
<td>2.1</td>
<td>–</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59912</td>
<td>86300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>45</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16881</td>
<td>57029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td>3.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
</tr>
<tr>
<td>social</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269504</td>
<td>1497135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203549046</td>
<td>213000000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783339</td>
<td>6716198</td>
<td>8.57</td>
<td>3.0/–</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>0.119</td>
<td>395</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1022</td>
<td>5103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460902</td>
<td>17000000</td>
<td>70.13</td>
<td>2.7</td>
<td>–</td>
<td>0.44</td>
<td>0.27</td>
<td>0.119</td>
<td>119, 157</td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10697</td>
<td>31992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4941</td>
<td>6594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>0.033</td>
<td>0.366</td>
<td>318</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1439</td>
<td>1723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1377</td>
<td>2213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24997</td>
<td>53248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>technological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2115</td>
<td>2240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices \( n \); total number of edges \( m \); mean degree \( z \); mean vertex-vertex distance \( \ell \); exponent \( \alpha \) of degree distribution if the distribution follows a power law (or “–” if not; in/out-degree exponents are given for directed graphs); clustering coefficient \( C^{(1)} \) from Eq. (3); clustering coefficient \( C^{(2)} \) from Eq. (6); and degree correlation coefficient \( r \), Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Small worlds in real networks

• For all real networks there are (on average) short paths between nodes of the network.
  – Largest path found in the IMDB actor network: 7

• Is this interesting?
  – Random graphs also have small diameter
    (d=\log n/\log \log n \text{ when } z=\omega(\log n))

• Short paths are not surprising and should be combined with other properties
  – ease of navigation
  – high clustering coefficient
Connected components

• For undirected graphs, the size and distribution of the connected components
  – is there a giant component that contains a large fraction of the nodes?
• Most known real undirected networks have a giant component
  – The giant component usually captures more than 80% of the nodes in the graph.
• For directed graphs, the size and distribution of strongly and weakly connected components
Connected components – definitions

- **Weakly connected components (WCC)**
  - Set of nodes such that from any node can go to any node via an undirected path

- **Strongly connected components (SCC)**
  - Set of nodes such that from any node can go to any node via a directed path.
  - **IN**: Nodes that can reach the SCC (but not in the SCC)
  - **OUT**: Nodes reachable by the SCC (but not in the SCC)
The bow-tie structure of the Web

The largest weakly connected component contains 90% of the nodes
SCC and WCC distribution

• The SCC and WCC sizes follow a power law distribution
  – the second largest SCC is significantly smaller
Web Cores

- **Cores**: Small complete bipartite graphs (of size 3x3, 4x3, 4x4)
  - Similar to the triangles for undirected graphs
- Found more frequently than expected on the Web graph
- Correspond to communities of enthusiasts (e.g., fans of Japanese rock bands)
Motifs

• Most networks have the same characteristics with respect to **global measurements**
  – can we say something about the **local structure** of the networks?

• **Motifs**: Find small subgraphs that are over-represented in the network
Example

- Motifs of size 3 in a directed graph
Finding interesting motifs

• Sample a part of the graph of size $S$
• Count the frequency of the motifs of interest
• Compare against the frequency of the motif in a random graph with the same number of nodes and the same degree distribution
Generating a random graph

- Find edges \((i,j)\) and \((x,y)\) such that edges \((i,y)\) and \((x,j)\) do not exist, and swap them
  - repeat for a large enough number of times

\[
\begin{align*}
G & \quad \text{G-swapped} \\
(i,j) & \quad (x,y) \\
(x,y) & \quad (i,j)
\end{align*}
\]

degrees of \(i,j,x,y\) are preserved
The feed-forward loop

• Over-represented in gene-regulation networks – a signal delay mechanism

Milo et al. 2002
Homophily

- Love of the same: People tend to have friends with common interests
  - Students separated by race and age
Measuring Homophily

If the fraction of cross-gender edges is significantly less than expected, then there is evidence for homophily.

Gender male with probability $p$ (fraction of males)
Gender female with probability $q$ (fraction of females)

Probability of cross-gender edge?

$$\frac{\text{#cross\_gender\_edges}}{\text{#edges}} \ll 2pq$$
Measuring Homophily

- We need to define what we mean by “significantly” less than

- Heterophily may also be interesting: In this case we want the ratio to be significantly larger than \(2pq\)

- Characteristics with more than two values:
  - Number of heterogeneous edges (edge between two nodes that are different)
Mechanisms Underlying Homophily: Selection and Social Influence

**Selection**: tendency of people to form friendships with others who are like them

**Socialization or Social Influence**: the existing social connections in a network are influencing the individual characteristics of the individuals

**Social Influence as the inverse of Selection**

Mutable & immutable characteristics
Longitudinal studies in which the social connections and the behaviors within a group are tracked over a period of time

Why?
- Study teenagers, scholastic achievements/drug use (peer pressure and selection)
- Relative impact?
- Effect of possible interventions (example, drug use)
The Interplay of Selection and Social Influence

Christakis and Fowler on obesity, 12,000 people over a period of 32-years

People more similar on obesity status to the network neighbors than if assigned randomly

Why?
(i) Because of selection effects, choose friends of similar obesity status,
(ii) Because of confounding effects of homophily according to other characteristics that correlate with obesity
(iii) Because changes in the obesity status of person’s friends was exerting an influence that affected her

(iii) As well -> “contagion” in a social sense
Tracking Link Formation in Online Data: interplay between selection and social influence

- Underlying social network
- Measure for behavioral similarity

Wikipedia

*Node*: Wikipedia editor who maintains a user account and user talk page

*Link*: if they have communicated with one writing on the user talk page of the other

Editor’s behavior: set of articles she has edited

**FACT**: Wikipedia editors who have communicated are significantly more similar in their behavior than pairs of Wikipedia editors who have not (homomophily), *why?*

Selection (editors form connections with those have edited the same articles) vs Social Influence (editors are led to the articles of people they talk to)
Tracking Link Formation in Online Data: interplay between selection and social influence

Actions in Wikipedia are time-stamped
For each pair of editors A and B who have ever communicated,
- Record their similarity over time
- Time 0 when they first communicated -- Time moves in discrete units, advancing by one “tick” whenever either A or B performs an action on Wikipedia
- Plot one curve for each pair of editors
Average, single plot: average level of similarity relative to the time of first interaction

Similarity is clearly increasing both before and after the moment of first interaction (both selection and social influence)
Not symmetric around time 0 (particular role on similarity): Significant increase before they meet
Blue line shows similarity of a random pair (non-interacting)
References

• S. N. Dorogovstev and J. F. F. Mendez, Evolution of Networks: From Biological Nets to the Internet and WWW.
NETWORK MODELS
What is a network model?

• Informally, a network model is a process (randomized or deterministic) for generating a graph of arbitrary size.

• Models of static graphs
  – input: a set of parameters $\Phi$, and the size of the graph $n$
  – output: a graph $G(\Phi,n)$

• Models of evolving graphs
  – input: a set of parameters $\Phi$, and an initial graph $G_0$
  – output: a graph $G_t$ for each time $t$
Families of random graphs

• A deterministic model $D$ defines a single graph for each value of $n$ (or $t$)

• A randomized model $R$ defines a probability space $\langle G_n, P \rangle$ where $G_n$ is the set of all graphs of size $n$, and $P$ a probability distribution over the set $G_n$ (similarly for $t$)
  – we call this a family of random graphs $R$, or a random graph $R$
Why do we care?

• Creating models for real-life graphs is important for several reasons
  – Create data for simulations of processes on networks
  – Identify the underlying mechanisms that govern the network generation
  – Predict the evolution of networks
Erdös-Renyi Random graphs

Paul Erdös (1913-1996)
Erdös-Renyi Random Graphs

• The $G_{n,p}$ model
  – input: the number of vertices $n$, and a parameter $p$, $0 \leq p \leq 1$
  – process: for each pair $(i,j)$, generate the edge $(i,j)$ independently with probability $p$

• Related, but not identical: The $G_{n,m}$ model
  – process: select $m$ edges uniformly at random
Random graphs degree distributions

- The degree distribution follows a **binomial**

\[ p(k) = \binom{n}{k} p^k (1-p)^{n-k} \]

- Assuming \( z=np \) is fixed, as \( n \to \infty \), \( B(n,k,p) \) is approximated by a **Poisson** distribution

\[ p(k) = P(k; z) = \frac{z^k}{k!} e^{-z} \]

- Highly concentrated around the mean, with a tail that drops **exponentially**
Other properties

• Clustering coefficient
  – \( C = p \)

• Diameter (maximum path)
  – \( L = \frac{\log n}{\log z} \)
The giant component

• Let $z=np$ be the average degree

• If $z < 1$, then almost surely, the largest component has size at most $O(\ln n)$

• If $z > 1$, then almost surely, the largest component has size $\Theta(n)$. The second largest component has size $O(\ln n)$

• If $z = \omega(\ln n)$, then the graph is almost surely connected.
The phase transition

• When $z=1$, there is a phase transition
  – The largest component is $O(n^{2/3})$
  – The sizes of the components follow a power-law distribution.
Phase transitions

- **Phase transitions** (a.k.a. Threshold Phenomena, Critical phenomena) are observed in a variety of natural or human processes, and they have been studied extensively by Physicists and Mathematicians
  - Also, in popular science: “The tipping point”

- **Examples**
  - Water becoming ice
  - Percolation
  - Giant components in graphs

- In all of these examples, the transition from one state to another (e.g., from water to ice) happens almost instantaneously when a parameter crosses a **threshold**

- At the threshold value we have **critical phenomena**, and the appearance of **Power Laws**
  - There is no characteristic scale.
Random graphs and real life

• A beautiful and elegant theory studied exhaustively

• Random graphs had been used as idealized network models

• Unfortunately, they don’t capture reality...
Departing from the ER model

• We need models that better capture the characteristics of real graphs
  – degree sequences
  – clustering coefficient
  – short paths
Graphs with given degree sequences

• The configuration model
  – input: the degree sequence \([d_1, d_2, \ldots, d_n]\)
  – process:
    • Create \(d_i\) copies of node \(i\)
    • Take a random matching (pairing) of the copies
      – self-loops and multiple edges are allowed

• Uniform distribution over the graphs with the given degree sequence
Example

- Suppose that the degree sequence is
  
  \[
  4 \quad 1 \quad 3 \quad 2
  \]

- Create multiple copies of the nodes

- Pair the nodes uniformly at random
- Generate the resulting network
Power-law graphs

• The critical value for the exponent $\alpha$ is

$$\alpha = 3.4788...$$

• The clustering coefficient is

$$C \propto n^{-\beta} \quad \beta = \frac{3\alpha - 7}{\alpha - 1}$$

• When $\alpha < 7/3$ the clustering coefficient increases with $n$
However...

• The problem is that these models are too contrived

• It would be more interesting if the network structure emerged as a side product of a stochastic process rather than fixing its properties in advance.
Preferential Attachment

• Power-laws are connected with the rich-get-richer process
  – Those who already have a lot, are more likely to receive more

• It can be shown mathematically that this will result in power-law distribution.
  – Explains power-laws in income distribution or city population

• In networks this is called “Preferential Attachment”:
  – Nodes with high degree are more likely to get more neighbors.
Preferential Attachment in Networks

• First considered by [Price 65] as a model for citation networks (directed)
  – each new paper is generated with $m$ citations (mean)
  – new papers cite previous papers with probability proportional to their in-degree (citations)
  – what about papers without any citations?
    • each paper is considered to have a “default” $a$ citations
    • probability of citing a paper with degree $k$, proportional to $k+a$

• Power law with exponent $\alpha = 2+a/m$
Barabasi-Albert model

• The BA model (undirected graph)
  – **input**: some initial subgraph $G_0$, and $m$ the number of edges per new node
  – **the process**:
    • nodes arrive one at the time
    • each node connects to $m$ other nodes selecting them with probability proportional to their degree
    • if $[d_1,\ldots,d_t]$ is the degree sequence at time $t$, the node $t+1$ links to node $i$ with probability
      \[
      \frac{d_i}{\sum_i d_i} = \frac{d_i}{2mt}
      \]
  • Results in power-law with exponent $\alpha = 3$
The mathematicians point of view
[Bollobas-Riordan]

- Self loops and multiple edges are allowed
- For the single edge problem:
  - At time $t$, a new vertex $v$, connects to an existing vertex $u$ with probability
    \[
    \frac{d_u}{2t - 1}
    \]
  - it creates a self-loop with probability
    \[
    \frac{1}{2t - 1}
    \]
- If $m$ edges, then they are inserted sequentially, as if inserting $m$ nodes
  - the problem reduces to studying the single edge problem.
Preferential attachment graphs

• Expected diameter
  – if \( m = 1 \), the diameter is \( \Theta(\log n) \)
  – if \( m > 1 \), the diameter is \( \Theta(\log n / \log \log n) \)

• Expected clustering coefficient is small

\[
E[C^{(2)}] = \frac{m - 1}{8} \frac{\log^2 n}{n}
\]
Weaknesses of the BA model

- **Technical issues:**
  - It is not directed (not good as a model for the Web) and when directed it gives acyclic graphs
  - It focuses mainly on the (in-) degree and does not take into account other parameters (out-degree distribution, components, clustering coefficient)
  - It correlates age with degree which is not always the case

- **Academic issues**
  - the model redisCOVERS the wheel
  - preferential attachment is not the answer to every power-law
  - what does “scale-free” mean exactly?

- **Yet, it was a breakthrough in the network research**
  - It reintroduced the preferential attachment process
  - It defined the notion of scale free graphs
  - It popularized Network Science
Variations of the BA model

- Many variations have been considered some in order to address the problems with the vanilla BA model
  - edge rewiring, appearance and disappearance
  - fitness parameters
  - variable mean degree
  - non-linear preferential attachment
    - surprisingly, only linear preferential attachment yields power-law graphs
Empirical observations for the Web graph

- In a large scale experimental study by Kumar et al, they observed that the Web contains a large number of small bipartite cliques (cores)
  - the topical structure of the Web

- Such subgraphs are highly unlikely in random graphs
- They are also unlikely in the BA model
- Can we create a model that will have high concentration of small cliques?
Copying model

• Input:
  – the out-degree $d$ (constant) of each node
  – a parameter $\alpha$

• The process:
  – Nodes arrive one at the time
  – A new node selects uniformly one of the existing nodes as a prototype
  – The new node creates $d$ outgoing links. For the $i^{th}$ link
    • with probability $\alpha$ it copies the $i$-th link of the prototype node
    • with probability $1-\alpha$ it selects the target of the link uniformly at random
An example

• $d = 3$
Copying model properties

• Power law degree distribution with exponent $\beta = (2-\alpha)/(1- \alpha)$

• Number of bipartite cliques of size $i \times d$ is $n e^{-i}$

• The model has also found applications in biological networks
  – copying mechanism in gene mutations
Small world Phenomena

• So far we focused on obtaining graphs with power-law distributions on the degrees. What about other properties?
  – **Clustering coefficient**: real-life networks tend to have high clustering coefficient
    • This property can appear if edges that close triangles have higher probability
  – **Short paths**: real-life networks are “small worlds”
    • This property can appear by adding random edges.
  – Can we combine these two properties?
Clustering Coefficient

• How can you create a graph with high clustering coefficient?

• High clustering coefficient but long paths
Small-world Graphs

• According to Watts [W99]
  – Large networks \((n \gg 1)\)
  – Sparse connectivity (avg degree \(z \ll n\))
  – No central node \((k_{\text{max}} \ll n)\)
  – Large clustering coefficient (larger than in random graphs of same size)
  – Short average paths (~\(\log n\), close to those of random graphs of the same size)
The Caveman Model [W99]

• The random graph
  – edges are generated completely at random
  – low avg. path length $L \leq \log n / \log z$
  – low clustering coefficient $C \sim \frac{z}{n}$

• The Caveman model
  – edges follow a structure
  – high avg. path length $L \sim \frac{n}{z}$
  – high clustering coefficient $C \sim 1 - O(1/z)$

• Can we interpolate between the two?
Mixing order with randomness

• Inspired by the work of Solmonoff and Rapoport
  – nodes that share neighbors should have higher probability to be connected
• Generate an edge between \( i \) and \( j \) with probability proportional to \( R_{ij} \)

\[
R_{ij} = \begin{cases} 
1 & \text{if } m_{ij} \geq z \\
\left( \frac{m_{ij}}{z} \right)^\alpha (1-p) + p & \text{if } 0 < m_{ij} < z \\
p & \text{if } m_{ij} = 0 
\end{cases}
\]

\( m_{ij} = \text{number of common neighbors of } i \text{ and } j \)

\( z = \text{average degree (high)} \)

\( p = \text{very small value} \)

• When \( \alpha = 0 \), edges are placed only between nodes with common neighbors (caveman model)
• When \( \alpha \to \infty \), edges are essentially independent of the common neighbors (except for rare cases)
• For intermediate values we obtain a combination of order and randomness
Algorithm

• Start with a ring
• For $i = 1 \ldots n$
  – Select a vertex $j$ with probability proportional to $R_{ij}$ and generate an edge $(i,j)$
• Repeat until $z$ edges are added to each vertex
Clustering coefficient – Avg path length

small world graphs
Watts and Strogatz model [WS98]

- Start with a ring, where every node is connected to the next \( z \) nodes.
- With probability \( p \), rewire every edge (or, add a shortcut) to a uniformly chosen destination.
  - Granovetter, “The strength of weak ties”

![Diagram](image-url)

- Order when \( p = 0 \)
- Randomness when \( p = 1 \)
- Intermediate states for \( 0 < p < 1 \)
Clustering Coefficient – Characteristic Path Length

When \( p = 0 \), \( C = \frac{3(k-2)}{4(k-1)} \sim \frac{3}{4} \)

\( L = \frac{n}{k} \)

For small \( p \), \( C \sim \frac{3}{4} \)

\( L \sim \log n \)
Graph Theory Results

• Graph theorist failed to be impressed. Adding random edges is known to decrease the diameter of a graph.
Network models and temporal evolution

• For most of the existing models it is assumed that
  – number of edges grows linearly with the number of nodes
  – the diameter grows at rate $\log n$, or $\log \log n$

• What about real graphs?
  – Leskovec, Kleinberg, Faloutsos 2005
Densification laws

- In real-life networks the average degree increases! – networks become denser!

\[ E(t) \propto N(t)^\alpha \]

\( \alpha = \) densification exponent

Graphs showing:
- Scientific citation network: \( E(t) \propto N(t)^{1.69} \)
- Internet: \( E(t) \propto N(t)^{1.18} \)
More examples

- The densification exponent $1 \leq \alpha \leq 2$
  - $\alpha = 1$: linear growth – constant out degree
  - $\alpha = 2$: quadratic growth - clique
What about diameter?

- **Effective diameter**: the interpolated value where 90% of node pairs are reachable
Diameter shrinks

scientific citation network

Internet

affiliation network

patent citation network
Modeling Densification

• Existing graph generation models do not capture the Densification Power Law and Shrinking diameters

• Can we find a simple model of local behavior, which naturally leads to observed phenomena?

• Two proposed models
  – Community Guided Attachment – obeys Densification
  – Forest Fire model – obeys Densification, Shrinking diameter (and Power Law degree distribution)
Community structure

- Let’s assume the community structure
- One expects many within-group friendships and fewer cross-group ones
- How hard is it to cross communities?

Self-similar university community structure
Fundamental Assumption

• The cross-community linking probability of nodes at tree-distance $h$ (the height of the least common ancestor) is **scale-free**

• We propose cross-community linking probability:

$$f(h) = c^{-h}$$

where: $c \geq 1$ ... the **Difficulty constant**

$h$ ... tree-distance
**Densification Power Law**

- **Theorem**: The Community Guided Attachment leads to Densification Power Law with exponent $a = 2 - \log_b(c)$

  $a$ ... densification exponent  \( E(t) \propto N(t)^a \)

  $b$ ... community structure branching factor

  $c$ ... difficulty constant
Theorem: \[ a = 2 - \log_b(c) \]

- Gives any non-integer Densification exponent
- If \( c = 1 \): easy to cross communities
  - Then: \( a = 2 \), quadratic growth of edges – near clique
- If \( c = b \): hard to cross communities
  - Then: \( a = 1 \), linear growth of edges – constant out-degree
Room for Improvement

• Community Guided Attachment explains Densification Power Law

• Issues:
  – Requires explicit Community structure
  – Does not obey Shrinking Diameters

• The ”Forrest Fire” model
“Forest Fire” model – Wish List

• We want:
  – no explicit Community structure
  – Shrinking diameters
  – and:
    • “Rich get richer” attachment process, to get heavy-tailed in-degrees
    • “Copying” model, to lead to communities
    • Community Guided Attachment, to produce Densification Power Law
“Forest Fire” model – Intuition

• How do authors identify references?
  1. Find first paper and cite it
  2. Follow a few citations, make citations
  3. Continue recursively
  4. From time to time use bibliographic tools (e.g. Google Scholar) and chase back-links
“Forest Fire” model – Intuition

• How do people make friends in a new environment?
  1. Find first a person and make friends
  2. From time to time get introduced to their friends
  3. Continue recursively

• Forest Fire model imitates exactly this process
“Forest Fire” – the Model

- A node arrives
- Randomly chooses an “ambassador”
- Starts burning nodes (with probability $p$) and adds links to burned nodes
- “Fire” spreads recursively
Forest Fire in Action (1)

- Forest Fire generates graphs that **Densify** and have **Shrinking Diameter**

![Graphs showing densification and diameter](image)

- Densification: $N(t) \propto 1.21$
- Diameter: $\text{diameter} = 0.83 \times 1.21^t, R^2 = 1.00$
Forest Fire in Action (2)

- Forest Fire also generates graphs with heavy-tailed degree distribution.
Forest Fire model – Justification

• Densification Power Law:
  – Similar to Community Guided Attachment
  – The probability of linking decays exponentially with the distance – Densification Power Law

• Power law out-degrees:
  – From time to time we get large fires

• Power law in-degrees:
  – The fire is more likely to reach hubs
Forest Fire model – Justification

• Communities:
  – Newcomer copies neighbors’ links

• Shrinking diameter
Kronecker graphs

- **Kronecker graphs** are a model for generating graphs using the *Kronecker product* matrix operation
  - Leskovec, Chakrabarti, Kleinberg, Faloutsos, PKDD 2005

- **Kronecker graphs** have *rich properties*:
  - **Static Patterns**
    - Power Law Degree Distribution
    - Small Diameter
    - Power Law Eigenvalue and Eigenvector Distribution
  - **Temporal Patterns**
    - Densification Power Law
    - Shrinking/Constant Diameter

- **Kronecker graphs** are *analytically tractable*
Idea: Recursive graph generation

- Intuition: self-similarity leads to power-laws
- Try to mimic recursive graph / community growth
- There are many obvious (but wrong) ways:
  - Kronecker Product is a way of generating self-similar matrices
Kronecker product: Graph

**Adjacency matrix**

\[
\begin{array}{ccc}
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1 \\
\end{array}
\]

(3x3)

**Intermediate stage**

**Adjacency matrix**

\[
\begin{array}{ccc}
G_1 & G_1 & 0 \\
G_1 & G_1 & G_1 \\
0 & G_1 & G_1 \\
\end{array}
\]

(9x9)

\[G_2 = G_1 \otimes G_1\]
The Kronecker product of matrices $A$ and $B$ is given by

$$C = A \otimes B \equiv \begin{pmatrix}
  a_{1,1}B & a_{1,2}B & \ldots & a_{1,m}B \\
  a_{2,1}B & a_{2,2}B & \ldots & a_{2,m}B \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{n,1}B & a_{n,2}B & \ldots & a_{n,m}B
\end{pmatrix}_{N*K x M*L}$$

We define a Kronecker product of two graphs as a Kronecker product of their adjacency matrices.
Kronecker graphs

• We create the self-similar graphs recursively
  – Start with an initiator graph $G_1$ on $N_1$ nodes and $E_1$ edges
  – The recursion will then produce larger graphs $G_2, G_3, \ldots G_k$ on $N_1^k$ nodes

• We obtain a growing sequence of graphs by iterating the Kronecker product

$$G_k = \underbrace{G_1 \otimes G_1 \otimes \ldots G_1}_{k \ times}$$
Kronecker product: Graph

- Continuing multiplying with $G_1$ we obtain $G_2$ and so on ...

$G_1$

\begin{align*}
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1 \\
\end{align*}

\begin{align*}
\text{adjacency matrix of } G_2 \\
\begin{bmatrix}
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
\end{bmatrix}
\end{align*}
Kronecker product: Graph

- Continuing multiplying with $G_1$ we obtain $G_2$ and so on ...

Each cell in $G_2$ is the product by two cells in $G_1$
Each cell in $G_3$ is the product of three cells in $G_1$ and so on

$G_2$ adjacency matrix
Example

(a) $K_3$ adjacency matrix (27 $\times$ 27)

(b) $K_4$ adjacency matrix (81 $\times$ 81)
Examples

Initiator $K_1$

$K_1$ adjacency matrix

$K_3$ adjacency matrix
Kronecker graphs: Intuition

- Recursive growth of graph communities
  - Nodes get expanded to micro communities
  - Nodes in sub-community link among themselves and to nodes from different communities as determined by the original graph $G_1$

$$G_2 = G_1 \otimes G_1$$
Kronecker graphs

- Kronecker graphs have nice properties but they are deterministic and the distributions we obtain are not smooth:

![Graph](image)

(a) Kronecker initiator $K_1$

(b) Degree distribution of $K_6$

(c) Network value of $K_6$

$6^{th}$ Kronecker power of $K_1$

Figure 5: The “staircase” effect. Kronecker initiator and the degree distribution and network value plot for the $6^{th}$ Kronecker power of the initiator. Notice the non-smoothness of the curves.
Stochastic Kronecker graphs

- Create $N_1 \times N_1$ probability matrix $\Theta_1$
- Compute the $k^{th}$ Kronecker power $\Theta_k$
- For each entry $p_{uv}$ of $\Theta_k$ include an edge $(u,v)$ with probability $p_{uv}$

$\Theta_1 = \begin{pmatrix} 0.5 & 0.2 \\ 0.1 & 0.3 \end{pmatrix}$

$\Theta_2 = \Theta_1 \otimes \Theta_1$

$\Theta_2 = \begin{pmatrix} 0.25 & 0.10 & 0.10 & 0.04 \\ 0.05 & 0.15 & 0.02 & 0.06 \\ 0.05 & 0.02 & 0.15 & 0.06 \\ 0.01 & 0.03 & 0.03 & 0.09 \end{pmatrix}$

Instance matrix $K_2$

For each $p_{uv}$ flip Bernoulli coin

Probability of edge $p_{uv}$
Stochastic Kronecker graphs: Intuition

- **Node attribute representation**
  - Nodes are described by $k$ features
    - [in Ioannina, student, computer science]
    - $u=[1,1,0], \ v=[1,1,1]$  
  - Parameter matrix gives the linking probability
    - $p(u,v) = 0.5 \times 0.5 \times 0.1 = 0.025$

<table>
<thead>
<tr>
<th>Both in Ioannina</th>
<th>Both students</th>
<th>One CS one not</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Matrix Diagram" /></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><img src="image" alt="Image" /></td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td><img src="image" alt="Image" /></td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

We could have different probabilities for different attributes
Kronecker graph construction

• We can construct the graph by flipping a coin for each of the possible edges.
  – But this is expensive, quadratic number of coins to flip.
• We can exploit the recursive/hierarchical nature of Kronecker graphs
Kronecker graph construction

• If for $P_1$ we have that $E_1 = \sum_{ij} \theta_{ij}$ then the number of edges is normally distributed with expectation $E_1^k$

• Process:
  – Sample the number of edges from the normal distribution
  – For each edge to be added, descend to the position of the edge:
    • Pick a top-level cell with probability $\theta_{ij}/E_1$
    • Within the top-level cell repeat recursively
    • Until you have gone down $k$ levels
To generate the edge \((v_2, v_3)\) first we pick the top quadrant and then within that we pick the exact cell of the matrix.
Properties of Kronecker graphs

• We prove that Kronecker multiplication generates graphs that obey [PKDD’05]
  – Properties of static networks
    ✓ Power Law Degree Distribution
    ✓ Power Law eigenvalue and eigenvector distribution
    ✓ Small Diameter
  – Properties of dynamic networks
    ✓ Densification Power Law
    ✓ Shrinking/Stabilizing Diameter

• Good news: Kronecker graphs have the necessary expressive power
Experiments

• Use a 4-star as the graph $G_1$

\[
\begin{array}{cccc}
1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 \\
\end{array}
\quad
\begin{array}{cccc}
\alpha & \alpha & \alpha & \alpha \\
\alpha & \alpha & \beta & \beta \\
\alpha & \beta & \alpha & \beta \\
\alpha & \beta & \beta & \alpha \\
\end{array}
\]

• Make the matrix **stochastic** by having probability $\alpha$ for all edges and $\beta$ for all non-edges in the matrix
Figure 7: Citation network (CIT-HEP-TH): Patterns from the real graph (top row), the deterministic Kronecker graph with $K_1$ being a star graph on 4 nodes (center + 3 satellites) (middle row), and the Stochastic Kronecker graph ($\alpha = 0.41, \beta = 0.11$ – bottom row). Static patterns: (a) is the PDF of degrees in the graph (log-log scale), and (b) the distribution of eigenvalues (log-log scale). Temporal patterns: (c) gives the effective diameter over time (linear-linear scale), and (d) is the number of edges versus number of nodes over time (log-log scale). Notice that the Stochastic Kronecker graphs qualitatively matches all the patterns very well.
Figure 8: *Autonomous systems (AS-ROUTEVIEWS)*: Real (top) versus Kronecker (bottom). Columns (a) and (b) show the degree distribution and the scree plot, as before. Columns (c) and (d) show two more static patterns (see text). Notice that, again, the Stochastic Kronecker graph matches well the properties of the real graph.
Figure 9: Effective diameter over time for a 4-node chain initiator graph. After each consecutive Kronecker power we measure the effective diameter. We use different settings of $\alpha$ parameter. $\alpha = 0.38, 0.43, 0.54$ and $\beta = 0$, respectively.
Threshold phenomena

(a) Largest component size
(b) Largest component size
(c) Effective diameter

Figure 10: Fraction of nodes in the largest weakly connected component ($N_c/N$) and the effective diameter for 4-star initiator graph. (a) We fix $\beta = 0.15$ and vary $\alpha$. (b) We vary both $\alpha$ and $\beta$. (c) Effective diameter of the network, if network is disconnected or very dense path lengths are short, the diameter is large when the network is barely connected.
Model estimation: approach

- How do we choose the parameters to match the properties of a real network?

- **Maximum likelihood estimation**
  - Given real graph $G$
  - Estimate Kronecker initiator graph $\Theta$ (e.g., \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}) which
  $$\arg \max_{\Theta} P(G | \Theta)$$

- We need to (efficiently) calculate
  $$P(G | \Theta)$$

- And maximize over $\Theta$ (e.g., using gradient descent)
Fitting Kronecker graphs

- Given a graph \( G \) and Kronecker matrix \( \Theta \) we calculate probability that \( \Theta \) generated \( G \) \( P(G|\Theta) \)

\[
P(G|\Theta) = \prod_{(u,v) \in G} \Theta_k[u,v] \prod_{(u,v) \notin G} (1 - \Theta_k[u,v])
\]
### Challenge 1: Node correspondence

#### Nodes are unlabeled

- Graphs $G'$ and $G''$ should have the same probability
  $$P(G' | \Theta) = P(G'' | \Theta)$$

#### One needs to consider all node correspondences $\sigma$

- All correspondences are a priori equally likely
  $$P(G | \Theta) = \sum_\sigma P(G | \Theta, \sigma)P(\sigma)$$

- There are $O(N!)$ correspondences

- Solution: Sample from the possible distributions

<table>
<thead>
<tr>
<th>$\Theta$</th>
<th>$\Theta_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

| $\sigma$ | $P(G' | \Theta)$ |
|-----------|------------------|
|           | 0.25 0.10 0.10 0.04 |
|           | 0.05 0.15 0.02 0.06 |
|           | 0.05 0.02 0.15 0.06 |
|           | 0.01 0.03 0.03 0.09 |

**Graphs $G'$ and $G''$:**
- $G'$: Nodes 1, 2, 3, 4
- $G''$: Nodes 2, 4, 1, 3

$$P(G' | \Theta) = P(G'' | \Theta)$$
Challenge 2: Calculating $P(G|\Theta,\sigma)$

- Calculating naively $P(G|\Theta,\sigma)$ takes $O(N^2)$
- Idea:
  - First calculate likelihood of empty graph, a graph with 0 edges
  - Correct the likelihood for edges that we observe in the graph
- By exploiting the structure of Kronecker product we obtain closed form for likelihood of an empty graph
Challenge 2: Calculating $P(G|\Theta, \sigma)$

- We approximate the likelihood:
  \[
  l(\Theta) \approx l_e(\Theta) + \sum_{(u,v) \in G} -\log(1 - \Theta_k[\sigma_u, \sigma_v]) + \log(\Theta_k[\sigma_u, \sigma_v])
  \]
  
  - The sum goes only over the edges
  - Evaluating $P(G|\Theta, \sigma)$ takes $O(E)$ time
  - Real graphs are sparse, $E << N^2$
Experiments: real networks

• Experimental setup:
  – Given real graph
  – Stochastic gradient descent from random initial point
  – Obtain estimated parameters
  – Generate synthetic graphs
  – Compare properties of both graphs
• We do not fit the properties themselves
• We fit the likelihood and then compare the graph properties
AS graph (N=6500, E=26500)

- Autonomous systems (internet)
- We search the space of \(\sim 10^{50,000}\) permutations
- Fitting takes 20 minutes
- AS graph is undirected and estimated parameter matrix is symmetric:

\[
\begin{array}{cc}
0.98 & 0.58 \\
0.58 & 0.06 \\
\end{array}
\]
AS: comparing graph properties

- Generate synthetic graph using estimated parameters
- Compare the properties of two graphs
AS: comparing graph properties

- Spectral properties of graph adjacency matrices

![Scree plot](image1.png)

![Network value](image2.png)
Epinions graph (N=76k, E=510k)

- We search the space of $\sim 10^{1,000,000}$ permutations
- Fitting takes 2 hours
- The structure of the estimated parameter gives insight into the structure of the graph

<table>
<thead>
<tr>
<th>log degree</th>
<th>Epinions</th>
<th>Kronecker</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^0</td>
<td>10^4</td>
<td>10^4</td>
</tr>
<tr>
<td>10^1</td>
<td>10^3</td>
<td>10^3</td>
</tr>
<tr>
<td>10^2</td>
<td>10^2</td>
<td>10^2</td>
</tr>
<tr>
<td>10^3</td>
<td>10^1</td>
<td>10^1</td>
</tr>
<tr>
<td>10^4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>log # of reachable pairs</th>
<th>Epinions</th>
<th>Kronecker</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^5</td>
<td>10^6</td>
<td>10^6</td>
</tr>
<tr>
<td>10^6</td>
<td>10^7</td>
<td>10^7</td>
</tr>
<tr>
<td>10^7</td>
<td>10^8</td>
<td>10^8</td>
</tr>
<tr>
<td>10^8</td>
<td>10^9</td>
<td>10^9</td>
</tr>
<tr>
<td>10^9</td>
<td>10^10</td>
<td>10^10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of hops</th>
<th>Epinions</th>
<th>Kronecker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10^5</td>
<td>10^5</td>
</tr>
<tr>
<td>2</td>
<td>10^6</td>
<td>10^6</td>
</tr>
<tr>
<td>3</td>
<td>10^7</td>
<td>10^7</td>
</tr>
<tr>
<td>4</td>
<td>10^8</td>
<td>10^8</td>
</tr>
<tr>
<td>5</td>
<td>10^9</td>
<td>10^9</td>
</tr>
<tr>
<td>6</td>
<td>10^10</td>
<td>10^10</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Epinions graph (N=76k, E=510k)
Scalability

- Fitting scales \textit{linearly} with the number of edges
Conclusion

• Kronecker Graph model has
  – provable properties
  – small number of parameters
• We developed scalable algorithms for fitting Kronecker Graphs
• We can efficiently search large space ($\sim 10^{1,000,000}$) of permutations
• Kronecker graphs fit well real networks using few parameters
• We match graph properties without a priori deciding on which ones to fit
References


Acknowledgements: Christos Faloutsos, Jon Kleinberg, Zoubin Gharamani, Pall Melsted, Alan Frieze, Larry Wasserman, Carlos Guestrin, Deepay Chakrabarti
Acknowledgements

• Many thanks to Jure Leskovec for his slides from the KDD 2005 paper and his slides on Kronecker graphs
References

- B. Bollobas, *Mathematical Results in Scale-Free random Graphs*